SURYA GROUP OF

INSTITUTIONSNAANMUDHALVAN

IBM-ARTIFICIALINTELLIGENCE

RAHULDASS S

422221104028

AI-Powered Spam Classifier

BuildingaSmarterAI-PoweredSpamClassifier

DepartmentofComputerEngineering

1 INTRODUCTION

The Internet has become an essential part of people's lives.On the Internet, people send and receive messages throughemail services such as Gmail, Outlook, AOL Mail, etc. As-suming that one or more advertising messages are sent to aperson every day, if this person is not aware of spam, he orshe will easily fall into the trap of this type of message . Themain purpose of spam is to advertise, damage the recipient'ssystem and steal important information from the recipient ofamessage.

Spamdetectionsystemsfocusmoreonthetextualcontentofamessage.Thesetypesofsystemsfirstreceiveatextualdataset(message set) as input and, after performing pre-processingoperations such as removing stop words, normalizing, stem-ming, etc., the message is processed using various naturallanguageprocessingandmachinelearningalgorithms.Usually, people who want to send spam messages collect acollection of users' emails from blogs, forums, then write atargetedletter(withtheaimofadvertisingorstealing)fora collection of emails. Collected are sent. Upon seeing thismessage, the recipient of the email may read it and send theirimportant information (bank card details, password, etc.) inresponse. Users ofthe United Nations,for example, havebeenvictimsofthistypeofattack.

If users ofemail serversrecognise a messageasspam, it isbetter not to click on any links or attachments .

Spammerssometimesaddunsubscribeorunsubscribelinkstoverifythat your email address is active; these types of links usuallysteal information, so users should not click on them.

Spammessagesaredifficulttostopbecausetheycanbe sentthroughbotnets,whicharenetworksofpre-infectedcomputersthatmakeitdifficulttotrackandstoptheoriginal spam.

In the proposed approach, a set of useful and spam messagesis first collected through Gmail, then text pre-processing isperformed on the textual content of the messages using thespaCytool.Finally,byusingthemachinelearningalgorithms Naive Bayes (NB), Decision Tree C45 and Multilayer Per-ceptron (MLP) in Python programming language, the spamdetectionprocesshasbeen done.

1.1 WhatisspaCy

spaCyisafreeandopen-sourcelibraryfornaturallan-guage processing (NLP) in Python. It provides various NLPcapabilities such as Named Entity Recognition (NER), Part-Of-Speech(POS)tagging,dependencyparsing,andwordvectors. spaCy is designed to make it easy to build systemsfor information extraction, text classification, and other NLPtasks. It can handle large volumes of text data and is widelyusedin datascience andmachine learning.

1.2 NaiveBayes(NB)

NaiveBayesisasimpleandfastclassificationalgorithmbased on Bayes' theorem. It is used for assigning class labelsto problem instances represented as vectors of feature values.NaiveBayesclassifiersareacollectionofclas-sificationalgorithms that can be used for various tasks such as textclassification,spamfiltering,andsentimentanalysis.Thealgorithmassumesthatthefeaturesareindependent,making it computationally efficient and easy to implement.Naive Bayes models are often suitable for high-dimensionaldatasets and can provide good results with limited trainingdata.

1.3 DecisionTreeC45[17]

A decision tree is a non-parametric supervised learning al-gorithmusedforbothclassificationandregressiontasks.It is a decision support tool that uses a tree-like model ofdecisions and their possible consequences, including chanceevent outcomes. A decision tree is a flowchart-like tree struc-ture consisting of nodes, branches, internal nodes, and leafnodes.Therootnoderepresentstheentirepopulationorsample.Theinternalnodesrepresentfeaturesorattributesof the data, while the branches represent the possible valuesof these features. The leaf nodes represent the outcome ortargetvariable.

1.4 MultilayerPerceptron(MLP)

A Multilayer Perceptron (MLP) [18] is a fully connected classof feedforward Artificial Neural Network (ANN). It consistsof three types of layers: the input layer, output layer, and oneormorehiddenlayerswithmanyneuronsstackedtogether.In contrasttothePerceptron,wheretheneuronmusthaveanactivationfunctionthat imposesathresholdlikeReLU or sigmoid, neurons in an MLP can use any arbitraryactivation function. An MLP is trained using backpropaga-tion, which adjusts the weights of the connections betweenneurons to minimize the error between predicted and actualoutputs.

Therefore,inthecontinuationofthisarticleandinsection2,a series of works done in the field of spam detection havebeen examined, then in section 3, the proposed approach ispresentedinfull.Section4examinestheresultsobtainedfromthe proposed approach, and finally, Section 5 discusses theconclusionsoftheproposedapproach.

2 RELATEDWORK

Alotofworkhasbeendonetodetectspam,mostofitusing natural language processing toolssuchas NLTKandSVMmachinelearningalgorithms.However,inrecentyears,various machine learning and deep learning algorithms havealso been used to detect spam. Mohammad et a presented a3-layer processing system for spam detection. In the firstlayer of the proposed system, emailsarefedintothesystemasaninputdataset.Inthe second layer, variousnaturallanguageprocessinglibrariessuchasNLTKare used to perform preprocessing and feature extraction. Finally,inthethirdlayer,theNaiveBayesmachinelearning algorithmis also used to detect spam or usefulness of incoming emails.Marsono et alhave presenteda hardwarearchitecture withthepurposeofpreventivemanagementtodetectspamorusefulemails.Thearchitectureprovidedbythemreceived117 millionemails as input every secondandafter perform-ingvarioustextpre-processingoperations,itwasusedtodetectspamorusefulnessofemailsusingSVMmachinelearningalgorithm.

Tang et al [19] presented an approach based on assigning atrust value to an IP address. By extracting IPs from a collec-tion of spam emails, they created a dataset of IPs and trustvalues.TheythenusedSVMandRandomForestmachine

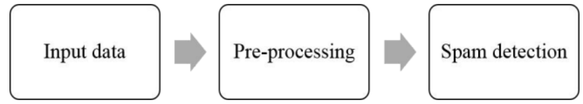

Fig.1.Implementationprocessoftheproposedapproach

learning algorithms to detect unreliable IPs and finally spam.The results showed better and faster accuracy of the SVMalgorithmthanthe RandomForest algorithm.

Yooetalpresentedaspamdetectionapproachbased onusergroupformation.Inthisapproach,byformingandcreatingdifferentgroupsofpeople,theusertrainsthesystemtorecognise theemailsreceivedfromthecreatedgroupsas useful emails and the restof the emails as spam. Thisapproach does not process the textual content of the messageor email. The proposed approach is implemented below. Inthe proposed approach,aset ofusefuland spammessagesis first collected through Gmail, then text preprocessing isperformed on the textual content of the messages using thespaCytool.Finally,byusingthemachinelearningalgorithms Naive Bayes (NB), Decision Tree C45 and Multilayer Percep-tron(MLP)inPythonprogramminglanguage,thespamdetection processhasbeen done.

3 THEPROPOSEDAPPROACH

To present the proposed approach, the author has used a3-step process. Figure 1 shows the process of the proposedapproach.AsshowninFigure1,itisclearthatinthefirststep,1500 messages or emails are collected from the Gmail serviceandprovidedasinputdatatothesecondstepoftheproposedsystem process. In the second step, various pre-processingoperations such as removal of stop words, removal of num-bers, normalisation and stemming are performed on the dataset collected from the previous step using the spaCy tool.Finally, in the third step, three algorithms Naive Bayes (NB),Decision Tree C45 and Multilayer Perceptron (MLP) are usedsimultaneouslytotrainthemodelandevaluatethemodelsin spamdetection .

3.1 Inputdata

In this step, 750 spam emails (from the spam folder) and 750useful emails (from the inbox folder), a total of 1500 emails,are extracted and collected as a primary dataset from theGmail account. The main reason for the fair division of theinput dataset into two parts, useful emails and spam, is tobalancetheresults ofthemodels builtinthethirdstep.

3.2 Pre-processing

Inthisstep,usingthespaCytool,variouspre-processingpro-cesses such as tokenization, removal of stop words, removalof numbers, normalization and stemming are performed ontheinputdatafromthefirststep.Thepurposeofpre-processing is to improve the quality of input data so that theprocess of detecting spam and useful messages can be donewithbetteraccuracyinthethirdstep.Forexample,Figure2

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
for token in doc:
    print(token.text)
```

0	1	2	3	4	5	6	7	8	9	10
Apple	is	looking	at	buying	U.K.	startup	for	$	1	billion

Fig.2.Tokenisationpre-processing

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
for token in doc:
    print(token.text, token.lemma_,token.is_stop)
```

TEXT	LEMMA	STOP
Apple	apple	False
is	be	True
looking	look	False
at	at	True
buying	buy	False
U.K.	u.k.	False
startup	startup	False
for	for	True
$	$	False
1	1	False
billion	billion	False

Fig.3.pre-processingoftokenization,stemmingandstop-worddetection

shows the tokenization pre-processing process on the contentof an email. As shown in Figure 2, the text of the messagecontains10words(tokens).Figure3showsthepre-processingoftokenization, stemmingand stopworddetection onthecontentofthe same email.

As shown in Figure 3, the TEXT column shows the words ofan email message, the LEMMA column shows the structuraland lexical root of each word in the TEXT column, and theSTOPcolumnshows whetherawordisa stop word.

3.3 Spamdetection

Oncethevariouspre-processingoperations(secondstep)have been performed on the input data and the quality datahas been obtained, different machine learning algorithms canbeusedto trainandevaluatethe results.

Therefore, in this step, 75% of the input data (1125 messages)are provided as training data to 3 algorithms Naive Bayes(NB), Decision Tree C45 and Multilayer Perceptron (MLP) tobuild 3 trained models. Finally, 3 trained models are evalu-atedontheremaining25%oftheinputdata(375messages)toperformtheprocessofspamdetectionandemailusefulness

		Predicted classes		
		TRUE	FALSE	
Actual classes	TRUE	135	53	188
	FALSE	10	177	187
		145	230	

TP	135
TN	177
FP	10
FN	53

Fig.4.ConfusionmatrixextractionfromthefinalNaiveBayes(NB)model

		Predicted classes		
		TRUE	FALSE	
Actual classes	TRUE	148	32	180
	FALSE	10	185	195
		158	217	

TP	148
TN	185
FP	10
FN	32

Fig.5.ConfusionmatrixextractionfromthefinalDecisionTreeC45model

to obtain 4 evaluation parameters: accuracy, precision, recallandf1-score.

4 OBSERVATIONS

Figures 5, 4 and 6 show the confusion matrices of each of themachine learning models Naive Bayes (NB), Decision TreeC45andMultilayerPerceptron(MLP)respectively,accordingto the real labels of the incoming emails and the predictedlabels. It shows by models. TRUE label means useful andFalse means an email is spam. Now, according to the valuesof the variables TP, FP, TN and FN of the confusion matricesofFigures5,4and6,itiseasytoobtaintheevaluationcriteriaofthe machinelearningalgorithms.

Figures 7, 8 and 9 show the statistical results of the accuracy,recall,precision,f1-scoreevaluationcriteriaoftheconstructedmodels Naive Bayes (NB), Decision Tree C45 and MultilayerPerceptron (MLP).As can be seen from Figures 8, 7 and 9,the model obtained from the Multilayer Perceptron (MLP)algorithm has accuracy (96%), recall (94%), precision (97%)andf1-score(96%).AlthoughtherecallofthemodelobtainedbytheDecisionTreeC45algorithmishigherthanthatofthe Naive Bayes (NB) and Multilayer Perceptron (MLP) algo-rithms.

Certainly, the preprocessing carried out with the spaCy natu-ral language processing library played a very important roleinthecriteriaobtained.

5 CONCLUSION

Spam or junk emails are a fundamental challenge that aresent to people's email accounts in bulk for the purpose ofadvertising,harmingandstealinginformationandfilling

		Predicted classes		
		TRUE	FALSE	
Actual classes	TRUE	163	10	173
	FALSE	5	197	202
		168	207	

TP	163
TN	197
FP	5
FN	10

Fig.6.ConfusionmatrixextractionfromthefinalMultilayerPerceptron(MLP)model

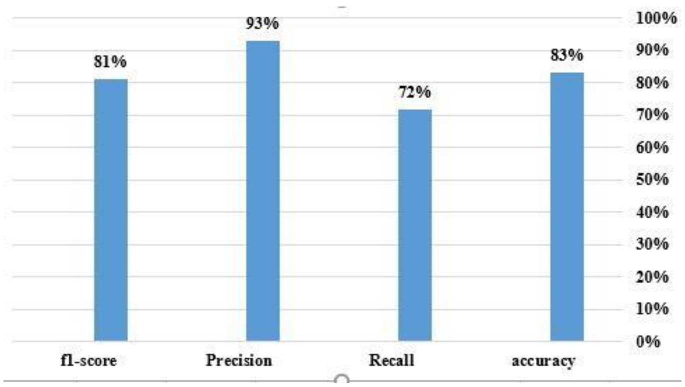

Fig.7.Statisticalresultsoftheaccuracy,recall,precisionandf1scoresofthefinalNaive Bayes(NB)model.

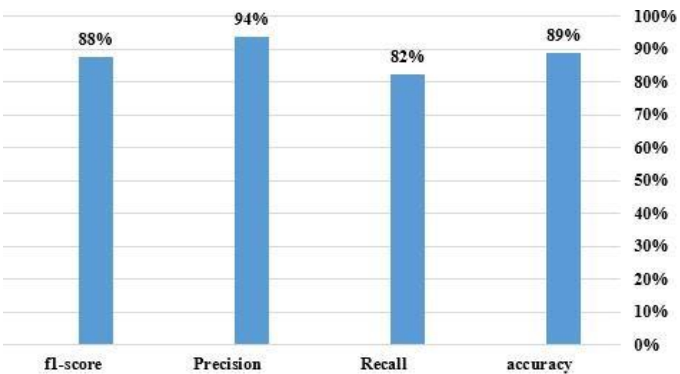

Fig.8.Statisticalresultsoftheaccuracy,recall,precisionandf1scoresofthefinalDecisionTree C45 model.

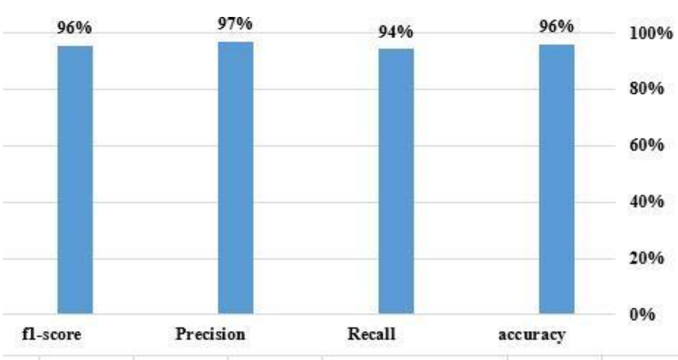

Fig.9.Statisticalresultsoftheaccuracy,recall,precisionandf1scoresofthefinalMultilayer Perceptron (MLP) model.

their inbox folders. In this article, the author first collectedand extracted 1500 useful and spam emails from the Gmailservice, then using the natural language processing libraryspaCy, he applied various text pre-processing operations onthetextcontentoftheemails,finallyusing3algorithmsMachine Learning Naive Bayes (NB), Decision Tree C45 andMultilayer Perceptron (MLP) in Python programming lan-guage was used to train and recognise spam emails collectedfromtheGmailservice.Observationsshowtheaccuracy (96%),recall(94%)andprecision(97%)oftheproposedapproachindetecting spam emails.

6 FUTUREWORKS

The use of machine learning techniques has been successfulin detectingand filteringspam. Inthe future, spamfiltersareexpectedtobecomemoreintelligentandabletodistinguish safe emails from those that need to be removedfromtheinbox.Anewmodelbasedondeeplearningalgorithms has been developed for automatic spam detectionandfiltering.Therefore,futureworkonspamdetection could include further research and development of machinelearning models that can accurately detect and filter spamwhileminimisingfalsepositives.Therecouldalsobeafocus on developing more intelligent filters that can adapt tonew types of spam attacks. It is also possible to identify andremovespamemailsusingtheMallettool[36]andtheDBPedia ontology.